D408–D414 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 26 November 2012*
doi:10.1093/nar/gks1177

DoBISCUIT: a database of secondary metabolite biosynthetic gene clusters

Natsuko Ichikawa, Machi Sasagawa, Mika Yamamoto, Hisayuki Komaki, Yumi Yoshida, Shuji Yamazaki and Nobuyuki Fujita*

Biological Resource Center, National Institute of Technology and Evaluation (NBRC), 2-49-10 Nishihara, Shibuya-ku, Tokyo 151-0006, Japan

Received August 13, 2012; Revised October 12, 2012; Accepted October 28, 2012

ABSTRACT

This article introduces DoBISCUIT (Database of BIoSynthesis clusters CUrated and InTegrated, http://www.bio.nite.go.jp/pks/), a literature-based, manually curated database of gene clusters for secondary metabolite biosynthesis. Bacterial secondary metabolites often show pharmacologically important activities and can serve as lead compounds and/or candidates for drug development. Biosynthesis of each secondary metabolite is catalyzed by a number of enzymes, usually encoded by a gene cluster. Although many scientific papers describe such gene clusters, the gene information is not always described in a comprehensive manner and the related information is rarely integrated. DoBISCUIT integrates the latest literature information and provides standardized gene/module/domain descriptions related to the gene clusters.

INTRODUCTION

Production of secondary metabolites is one of the industrially important features of bacteria, such as actinomycetes and myxobacteria. Various secondary metabolites (or their derivatives) produced by bacteria have been developed as antibiotics, antitumor drugs and immunosuppressive drugs (1). Therefore, bacterial secondary metabolites have an important role in the development of novel medicines.

Secondary metabolites usually comprise various chemical moieties, such as polyketide backbones, amino acid derivatives and sugars. Many enzymes are involved in the synthesis of secondary metabolites. Polyketide synthase (PKS) and non-ribosomal peptide synthetase (NRPS), which catalyze the elongation of polyketides and synthesis of oligopeptides, respectively, are the major enzymes of secondary metabolite synthesis. Enzymes responsible for the synthesis of other constitutive compounds, such as sugars, are often encoded by genes adjacent to PKS and NRPS genes. Through further tailoring events, such as glycosylation, alkylation and oxidation, structurally diverse and complex metabolites are finally synthesized. In addition, the production and transportation of secondary metabolites are strictly regulated by transcriptional regulators and transporters (2). Genes encoding such tailoring enzymes, transcriptional regulators and transporters are often located adjacent to PKS and NRPS genes. As a result, the whole set of genes responsible for the biosynthesis of each secondary metabolite are encoded in a large gene cluster spanning 10–100 kb.

Studies on secondary metabolite biosynthesis have mainly focused on PKS and NRPS because of their major roles in constructing complex carbon frameworks. A complex carbon structure is assembled sequentially from simple carbon building blocks, such as acyl-CoA and amino acids. The extension of each carbon unit is catalyzed by a set of functional domains, collectively termed as a 'module', encoded in a PKS and a NRPS. A minimal set of domains that function as a 'module' in a PKS generally comprises ketosynthase (KS), acyltransferase (AT) and acyl carrier protein (ACP) domains. The chemical structure of each starter/extender carbon unit can be predicted by examining substrate specificity determining residues of the AT domain and the presence of optional ketoreductase (KR), dehydratase (DH) and enoylreductase (ER) domains (3–6). Similarly, a minimal set of domains that function as a module of an NRPS generally comprises condensation (C), adenylation (A) and peptidyl carrier protein (PCP) domains. The specificity for each starter/extender amino acid is determined by active residues in the A domain, and loaded amino acids are modified by optional domains, such as methyltransferase, epimeration and reductase domains (7). Therefore, precise identification of functional domains and assignment of substrate specificity aid in determining the biosynthetic mechanism.

*To whom correspondence should be addressed. Tel: +81 3 3481 1972; Fax: +81 3 3481 8424; Email: fujita-nobuyuki@nite.go.jp

Since the first elucidation of the PKS gene cluster for erythromycin in the early 1990s (8–10), many gene clusters responsible for the biosynthesis of polyketide compounds have been reported and deposited in International Nucleotide Sequence Database Collection (INSDC) entries (DDBJ/GenBank/EMBL) (11,12). However, newer analysis data on previously uncharacterized genes or additional sequence data that extend the cluster region are rarely incorporated into the existing INSDC entries and are not stored in a comprehensive manner. Scientists must read and search a chain of reference citations to identify INSDC entries related to the biosynthetic cluster and to accumulate related knowledge.

Here, we developed a database of biosynthetic clusters of secondary metabolites called DoBISCUIT (Database of BIoSynthesis clusters CUrated and InTegrated). Currently, 72 known polyketide biosynthetic clusters are registered. Each biosynthetic cluster is manually curated in terms of sequence collection, reference collection and gene annotation. The database comprises cluster information pages with related references and gene description pages. Our database enables easy access to comprehensive information related to biosynthetic clusters and also serves as a useful reference to facilitate research and development related to secondary metabolites.

DATABASE CONTENT

Data sources and sequence collection

The current version of DoBISCUIT focuses on secondary metabolites derived from bacteria, especially from actinomycetes. The core data of DoBISCUIT are based on INSDC entries describing each biosynthetic cluster of a known bacterial secondary metabolite. Data collection started from a comprehensive review of the literature that reported discoveries of biosynthetic clusters. Articles were collected from PubMed using the search term 'biosynthesis cluster'. The corresponding INSDC accession numbers were extracted from these articles or by searching GenBank using the name of each compound. Further literature collection was achieved using a paper recommendation system, PubMedScan (http://medals.jp/pubmedscan/), which automatically reports articles highly related to a collection of literature.

In some cases, subsequent investigation revealed the existence of additional functional genes adjacent to a previously identified gene cluster. As a result, the complete biosynthetic cluster was divided into multiple INSDC entries, registered separately. For instance, the entire biosynthetic cluster of jadomycin B comprised five INSDC entries. The positional relationships between previously identified genes and newly found genes are described only in the original article and not assigned in each INSDC entry. In such cases, we collected all INSDC entries comprising the biosynthesis cluster (as far as we could determine) and reconstructed the whole cluster in DoBISCUIT. Currently, the database contains 72 gene clusters comprising 119 INSDC entries (Table 1).

Reference collection

Usually, three major phases are required in the investigation of a biosynthetic cluster: (i) identification of bioactive secondary metabolites, (ii) sequencing of the corresponding biosynthetic gene cluster and (iii) functional characterization of each gene in the cluster. This process can, therefore, produce multiple research articles. For example, there are at least 37 research articles related to the biosynthetic cluster of actinorhodin. In DoBISCUIT, we collected as many references as possible by extensive searching of PubMed using the compound name, organism name and gene name as search terms. Currently, we have collected 516 references for the 72 biosynthesis gene clusters (Table 1).

Gene descriptions

Gene descriptions in each INSDC entry vary considerably, depending on the submitters. Frequently, a gene product is represented by an abbreviation, such as capitalized gene symbol. Such a description gives little information about the function of the gene. In other cases, different submitters use different vocabulary to describe the same gene product. To develop a useful database, it is important to provide intelligible information for users. In DoBISCUIT, genes are described using a controlled vocabulary after manual curation, based on experimental data in reference articles and similarities to known proteins. If further analysis results are published after the release of biosynthetic cluster sequences, the gene descriptions are updated to represent their latest known function. Reference information is assigned to each gene to provide evidence for the annotation.

During the review of the biosynthetic clusters, we encountered cases where the function of a formerly uncharacterized gene could be inferred from similarities to characterized genes found in other gene clusters. In these cases, we annotated the gene based on the function of similar genes after careful evaluation of the level of similarity, e.g. having $>30\%$ amino acid identity with an experimentally confirmed protein. Of 2006 genes assigned in INSDC entries, we updated 1621 gene descriptions (Table 1).

According to our updated annotations, genes from biosynthetic clusters are expected to play various biological roles, including previously unrecognized roles. We classified all genes based on their biological role. The functional distribution of genes present in biosynthetic clusters is shown in Table 2.

USER AVAILABILITY AND WEB INTERFACE

The main content of DoBISCUIT comprises information pages about the biosynthetic cluster (Cluster information page) and each gene encoded in the biosynthetic cluster (CDS information page). Users can also search the contents of DoBISCUIT by keywords, module structures and sequence similarities.

On the home page, users are presented with the main table of biosynthetic clusters, with multiple options for viewing their attributes. By selecting one of the biosynthetic clusters listed in the table, users can view a Cluster information page.

Table 1. Number of sequences and references registered in DoBISCUIT[a]

Data type	Number
Gene clusters	72
Collected INSDC sequences	119
Collected references	516
Assigned genes	2006
Description changed from original INSDC entry	1621
Description accepted	196
Description not concerned[b]	189

[a]Based on the latest database release as of 3 October 2012.
[b]Genes proved not be involved in biosynthesis process.

Table 2. Functional categories and number of classified genes

Functional category	Number of CDS
Aglycon biosynthesis	
PKS	310
NRPS	26
PKS/NRPS hybrid	5
Other	54
Biosynthesis, modification and addition of aglycon units and moiety	
Extender unit	80
Starter unit	53
Sugar unit	251
Modification	
Hydroxylation	5
Methylation	38
Reduction	153
Other modification	111
Other function	
Transcriptional regulator	141
Translation	3
Transport	81
Resistance	12
Electron carrier	10
Biosynthesis of butyrolactone	4
Putative and unknown function	
General function prediction	191
Function unknown	45
Hypothetical protein	227

Cluster information page

This page shows integrated information about the biosynthesis cluster (Figure 1A) and has six sections: compound, original source, genomic map, PKS/NRPS modules, references and data download. The compound section displays the chemical structure, biological activities and various structural attributes of the secondary metabolite, such as chain length and sugar attachment. The original source section displays the bacterial strain from which the biosynthesis cluster sequences were obtained. Users can follow a hyperlink to access the culture collection distributing the strain. The original INSDC entries of the biosynthetic cluster are also displayed in this section. The genomic map section displays the coordinates of the genes in the biosynthetic cluster. If the biosynthetic cluster is represented by multiple INSDC entries, they are merged into a single map and the relative location of each entry is displayed on the map. Each gene is colored based on its biological function. The PKS/NRPS modules section displays the domain organization of each module in these enzymes. The deduced substrate of each AT or A domain is shown in the right-most column. Inactive domains are shown in lowercase letters. The reference section displays collected references concerning the biosynthetic cluster, with hyperlinks to PubMed records. The data download section allows users to download certain types of data files: nucleotide sequence of the cluster, CDS nucleotide/amino acid sequences in multi-FASTA format and curated annotations in CSV format or GenBank format.

A list of CDSs encoded by the biosynthetic cluster is displayed in another tab of the Cluster information page (Figure 1B). CDSs are ordered based on the relative position in the biosynthesis cluster. The list includes a summary of the annotation, including product name, gene name, keyword and functional category of each gene.

CDS information page

CDS information page, which can be accessed by selecting each CDS ID listed under the CDS list tag, shows integrated information about each CDS in the biosynthetic cluster (Figure 2). This page has six sections: location, annotation, genomic map, PKS/NRPS modules, sequence and features. The location section displays basic information about the CDS, such as position, length, source organism and INSDC entry. The annotation section displays functional information. Functional category, product name (in controlled vocabulary) and other notes assigned by the annotators are displayed. The original product name and gene name assigned in INSDC entries are also displayed side by side. References and corresponding UniProt entries are presented as the evidence of the annotation. If the gene was annotated based on similarities to other sequences, identifiers and hyperlinks for these similar sequences are displayed. The sequence section displays the nucleotide and amino acid sequences of the CDS. The displayed sequence can be switched between nucleic acid and protein using tab buttons. In the case of PKS/NRPS, each domain region is highlighted by a different color. Signature sequences of AT and A domains are also highlighted, with their respective substrates displayed in balloons (Figure 2). The feature section displays the results of automatic searches using bioinformatics tools. The 'Show BLAST table' button has a hyperlink to the result of a similarity search (BLASTP) (13) executed against the UniProt database (14). Domain assignments obtained by InterProScan (15) are also displayed.

Search menus

Various search menus are provided in DoBISCUIT. A simple text search form is provided in the upper right corner of all pages and other search menus can be accessed by following the links in the upper left panel.

In the simple text search, the search target is restricted to frequently used fields, i.e. compound name, organism name, product name and gene name. Search results are presented separately under 'Cluster' and 'CDS' tabs.

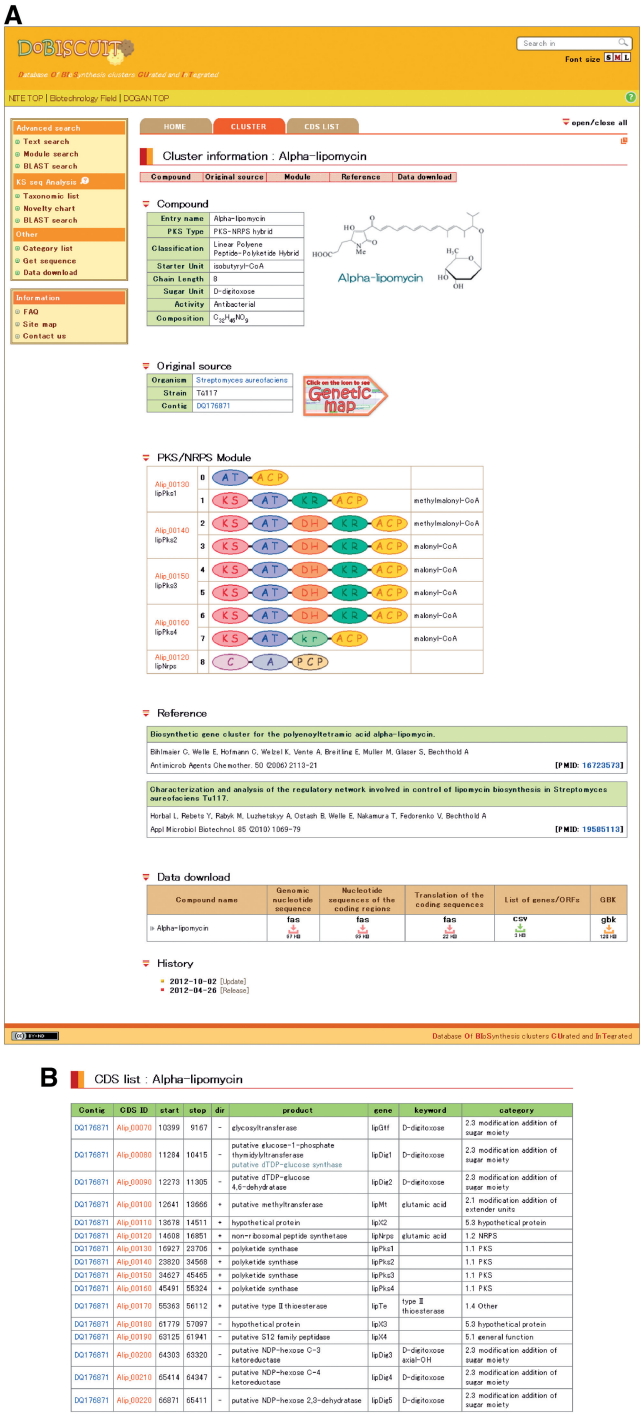

Figure 1. (**A**) Default view of the cluster information page for the alpha-lipomycin biosynthetic gene cluster. Information related to chemical compound, producing organism, gene coordinates, domain/modules, related references and downloadable flat files are displayed. (**B**) CDS list view of the alpha-lipomycin biosynthetic cluster. The list page is reached via the cluster information page by clicking 'CDS list' tab. Relative coordinate of CDSs in the biosynthetic cluster, annotated gene descriptions and functional categories are listed.

In the text search menu, users can execute more detailed searches within DoBISCUIT by entering search keywords, specifying the target fields and selecting the target clusters. Target gene clusters can be selected by their attributes, such as PKS type, attached sugar and chain length. Spaces between words are regarded as an 'AND' search term. The search result is displayed as a list of clusters matching the search conditions. The hyperlink can be followed to view a particular biosynthesis cluster page, or compound(s) of interest can be selected. Pressing the CDS tab permits browsing of the CDSs.

We also provide a module search menu to find PKSs and NRPSs containing a particular domain composition within the modules. All of the module patterns registered in DoBISCUIT are displayed in the upper part of the menu. Alternatively, auxiliary input boxes in the middle part of the menu can be used to specify the composition. The result of a module search displays a list of CDSs containing the entered domain composition.

To search homologous CDSs in DoBISCUIT, a BLAST utility is also provided. We provide several kinds of BLAST databases: cluster (containing the whole cluster sequences), CDS (containing all assigned CDSs) and domain (containing all biosynthesis-related domains assigned in CDSs). The BLAST search results are displayed separately as a list (top) and as an alignment (bottom). Clicking the 'B' button in the list part displays the alignment calculated by the bl2seq program (13,16) and clicking the 'T' button displays the alignment calculated by the T-COFFEE program (17).

DISCUSSION

The use of DoBISCUIT in genome mining

The number of genome projects is growing rapidly because of advances in sequencing technologies and decreasing costs. As of summer 2012, 103 genome projects intended for the genus *Streptomyces* are registered in the Genomes OnLine Database (GOLD) (18). Perhaps many of these genome projects intend to discover or investigate secondary metabolites produced by *Streptomyces* bacteria (19–21). Effective *in silico* identification of biosynthetic clusters from genome sequences is thought to be essential, and some useful web tools have been published (22–25). These web tools identify domains in PKS/NRPS proteins and propose similar known biosynthetic clusters to their own. However, in the next stage of genome mining, users will discover that the information cannot be obtained efficiently from suggested INSDC entries. DoBISCUIT can provide functional annotation of each gene and a comprehensive collection of references. Using a module search, users can obtain a list of CDSs containing the same domain composition as their own. Researchers will find it more appropriate to use a finely curated, concise database as a reference than searching a vast amount of patchy information.

The use of DoBISCUIT in combinatorial biosynthesis

Combinatorial biosynthesis approaches have been attracting attention for the generation of novel natural products and for the production of non-natural derivatives (26,27). Using recently developed gene manipulation technology, heterologous expression of biosynthetic clusters has been established in *Escherichia coli* (28–30) and *Streptomyces*

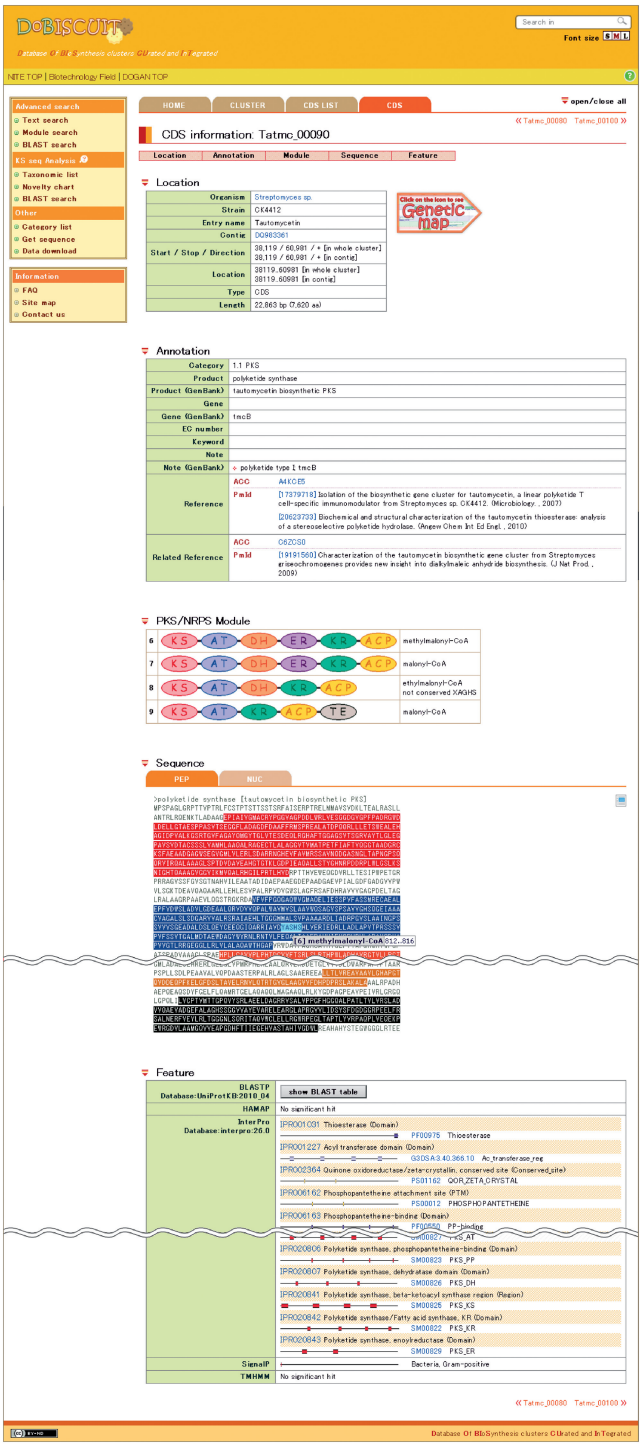

Figure 2. Default view of the CDS information page for a gene in the tautomycetin biosynthetic cluster. A hyperlink to the CDS information page is provided on the cluster information page and the CDS list page. Manual curation results, such as annotated product name, original product name, gene name and functional category are displayed in the annotation section. Experimental evidence for this gene is also provided in the reference column of the annotation section. Functionally important domains for formation of the polyketide chain are displayed in the PKS/NRPS module section. In the sequence section, different domains are indicated by different colors.

avermitilis (31). Genetic modification of biosynthetic clusters and/or introduction of a particular mutation also offer opportunities to obtain derivatives of original metabolites. DoBISCUIT can provide functionally classified lists of known biosynthetic cluster genes, which will enable users to easily identify genes encoding enzymes with appropriate specific activities as candidates for modification of the biosynthetic process. The CDS information page also provides detailed information on what reaction each gene product catalyzes. Users will be able to judge the potential applicability of genes to their combinatorial biosynthesis project.

The use of DoBISCUIT in assessing the novelty of biosynthetic genes

DoBISCUIT provides a curated set of domain sequences of known biosynthesis cluster enzymes as a BLAST database, allowing users to judge the novelty of their sequences using BLAST searches. However, only limited numbers of PKS and NRPS genes corresponding to known bioactive compounds have been identified so far. KS and A domains are essential constituents in PKSs and NRPSs, respectively, and the phylogenetic relationships of their sequences are closely related to the chemical structures of final products; therefore, sequencing analysis of these domains has been frequently used for assessing the potential of microorganisms to produce novel secondary metabolites (32,33). In such studies, PCR amplification of KS and A domains, followed by cloning and sequencing, is conventionally used, and the novelty of each domain is assessed by similarity to known domain sequences. The application of tagged amplicon sequencing on massively parallel sequencers can dramatically increase the throughput of such analysis. We massively sequenced the KS and A domains of 464 type strains of the genus *Streptomyces* and 333 antibiotic-producing actinomycetes preserved at NBRC (Komaki *et al.*, manuscript in preparation). Currently, the resultant nucleotide and amino acid sequences of >18 000 domains are available in the 'KS seq Analysis' menu of DoBISCUIT. BLAST searches against this data set ('BLAST search' in this menu) will complement the above approach for assessing the novelty of biosynthetic genes in user-collected strains, although each domain sequence has not yet been linked to its metabolite. The 'Novelty chart' in the same menu shows a graphical representation of the numbers and BLAST identities (against GenBank database) of independent (non-redundant) KS/A domain sequences assigned in each strain as indices of abundance and novelty, respectively (Figure 3). The 'Taxonomic list' in the menu allows users to access domain sequence data of all 797 strains according to their taxonomic names. These functions in the 'KS seq Analysis' menu will help users to select actinomycete strains suitable for their research purposes, based on the novelty and abundance of PKS/NRPS genes in each strain.

FUTURE PERSPECTIVES

DoBISCUIT (http://www.bio.nite.go.jp/pks/) enables easy access to comprehensive information related to

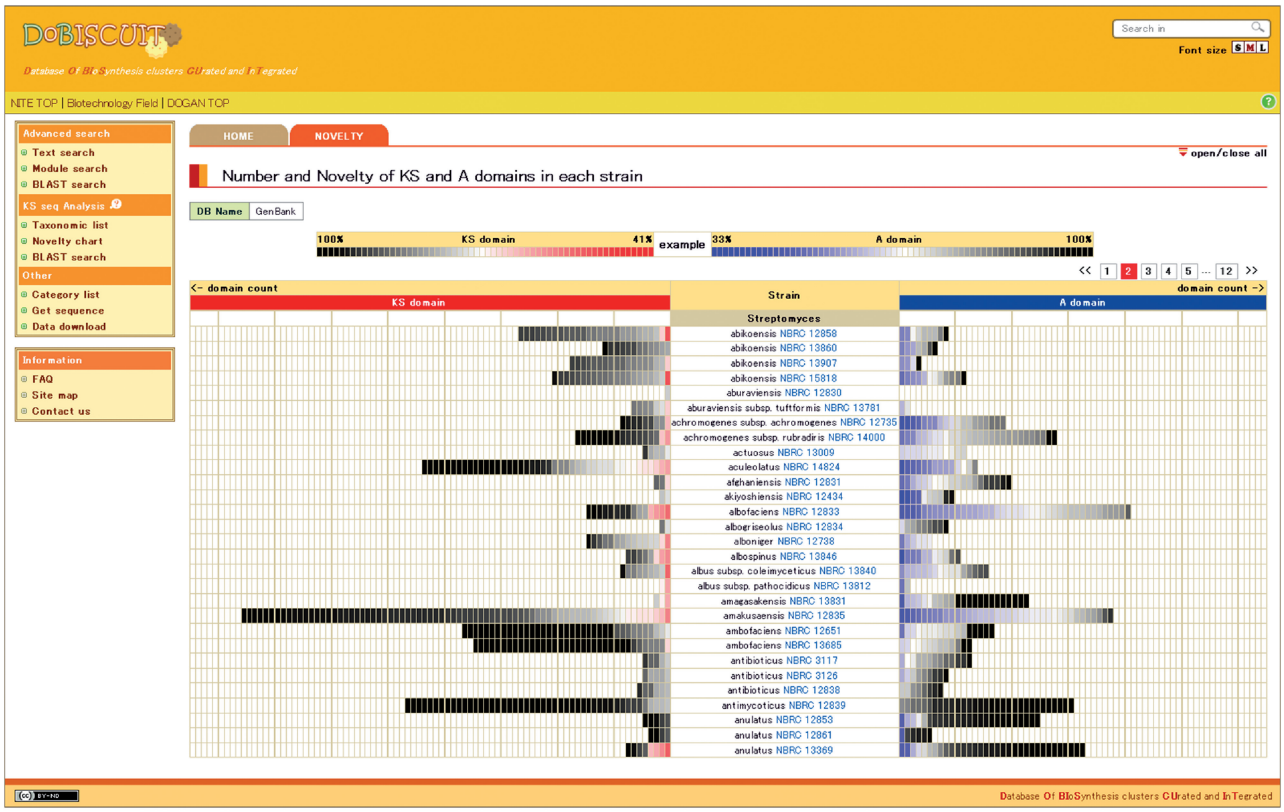

Figure 3. Novelty chart view of the KS domain and the A domain. Number and similarity to known sequences of KS (left) and A (right) domain sequences amplified from NBRC strains are displayed as a bar chart. The width of the horizontal rectangle represents the number, and the color gradient of each rectangle represents the degree of similarity for each NBRC strain shown in the center.

biosynthesis clusters and forms a standard reference for their investigation. The content of DoBISCUIT will be updated with new biosynthetic clusters and new findings for existing genes. We have already prepared information on 30 more biosynthetic clusters, which will be released soon. Although the current version of DoBISCUIT mainly focuses on PKS and NRPS in bacteria, especially in actinomycetes, secondary metabolites also comprise other compounds, such as thiopeptides, aminoglycosides and terpenoids, and are also found in other organisms, such as filamentous fungi. We aim to collect biosynthetic clusters representing a wider range of organisms and compounds in future versions of DoBISCUIT.

Our experience of curating a number of biosynthetic clusters allowed us to predict novel functions of previously uncharacterized genes located within clusters. For example, some uncharacterized genes within the chalcomycin, megalomicin and pikromycin clusters were predicted to encode helper proteins of glycosyltransferases. Further accumulation of cluster information and associated knowledge may help to understand the functions of hitherto unclassified genes.

ACKNOWLEDGEMENTS

We thank the members of the NBRC for their support, useful discussions and comments. We also thank Dr Sachiko Narita-Yamada for the annotations of type-I PKS biosynthetic clusters, and Dr. Kazuo Shin-ya for inspiring us to develop this database.

NOTE

In the latest version of DoBISCUIT, sequence alignments in BLAST menus are calculated only by the bl2seq program, but not by the T-COFFEE program, for simplicity and better performance.

FUNDING

Funding for open access charge: Ministry of Economy, Trade and Industry of Japan.

Conflict of interest statement. None declared.

REFERENCES

1. Newman,D.J. and Cragg,G.M. (2012) Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J. Nat. Prod.*, **75**, 311–335.
2. Park,S.R., Yoo,Y.J., Ban,Y.H. and Yoon,Y.J. (2010) Biosynthesis of rapamycin and its regulation: past achievements and recent progress. *J. Antibiot.*, **63**, 434–441.
3. Jenke-Kodama,H., Borner,T. and Dittmann,E. (2006) Natural biocombinatorics in the polyketide synthase genes of the actinobacterium Streptomyces avermitilis. *PLoS Comput. Biol.*, **2**, e132.

4. Khosla,C., Kapur,S. and Cane,D.E. (2009) Revisiting the modularity of modular polyketide synthases. *Curr. Opin. Chem. Biol.*, **13**, 135–143.
5. Del Vecchio,F., Petkovic,H., Kendrew,S.G., Low,L., Wilkinson,B., Lill,R., Cortes,J., Rudd,B.A., Staunton,J. and Leadlay,P.F. (2003) Active-site residue, domain and module swaps in modular polyketide synthases. *J. Ind. Microbiol. Biotechnol.*, **30**, 489–494.
6. Kakavas,S.J., Katz,L. and Stassi,D. (1997) Identification and characterization of the niddamycin polyketide synthase genes from Streptomyces caelestis. *J. Bacteriol.*, **179**, 7515–7522.
7. Stachelhaus,T., Mootz,H.D. and Marahiel,M.A. (1999) The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.*, **6**, 493–505.
8. Cortes,J., Haydock,S.F., Roberts,G.A., Bevitt,D.J. and Leadlay,P.F. (1990) An unusually large multifunctional polypeptide in the erythromycin-producing polyketide synthase of Saccharopolyspora erythraea. *Nature*, **348**, 176–178.
9. Donadio,S., Staver,M.J., McAlpine,J.B., Swanson,S.J. and Katz,L. (1991) Modular organization of genes required for complex polyketide biosynthesis. *Science*, **252**, 675–679.
10. Summers,R.G., Donadio,S., Staver,M.J., Wendt-Pienkowski,E., Hutchinson,C.R. and Katz,L. (1997) Sequencing and mutagenesis of genes from the erythromycin biosynthetic gene cluster of Saccharopolyspora erythraea that are involved in L-mycarose and D-desosamine production. *Microbiology*, **143 (Pt 10)**, 3251–3262.
11. Kodama,Y., Mashima,J., Kaminuma,E., Gojobori,T., Ogasawara,O., Takagi,T., Okubo,K. and Nakamura,Y. (2012) The DNA Data Bank of Japan launches a new resource, the DDBJ Omics Archive of functional genomics experiments. *Nucleic Acids Res.*, **40**, D38–D42.
12. Benson,D.A., Karsch-Mizrachi,I., Clark,K., Lipman,D.J., Ostell,J. and Sayers,E.W. (2012) GenBank. *Nucleic Acids Res.*, **40**, D48–D53.
13. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
14. UniProt Consortium. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.
15. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daugherty,L., Duquenne,L. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.
16. Tatusova,T.A. and Madden,T.L. (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.*, **174**, 247–250.
17. Di Tommaso,P., Moretti,S., Xenarios,I., Orobitg,M., Montanyola,A., Chang,J.M., Taly,J.F. and Notredame,C. (2011) T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res.*, **39**, W13–W17.
18. Pagani,I., Liolios,K., Jansson,J., Chen,I.M., Smirnova,T., Nosrat,B., Markowitz,V.M. and Kyrpides,N.C. (2012) The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.*, **40**, D571–D579.
19. Doroghazi,J.R., Ju,K.S., Brown,D.W., Labeda,D.P., Deng,Z., Metcalf,W.W., Chen,W. and Price,N.P. (2011) Genome sequences of three tunicamycin-producing Streptomyces strains, S. chartreusis NRRL 12338, S. chartreusis NRRL 3882, and S. lysosuperificus ATCC 31396. *J. Bacteriol.*, **193**, 7021–7022.
20. Qin,S., Zhang,H., Li,F., Zhu,B. and Zheng,H. (2012) Draft genome sequence of marine Streptomyces sp. strain W007, which produces angucyclinone antibiotics with a benz[a]anthracene skeleton. *J. Bacteriol.*, **194**, 1628–1629.
21. Klassen,J.L., Adams,S.M., Bramhacharya,S., Giles,S.S., Goodwin,L.A., Woyke,T. and Currie,C.R. (2011) Draft genome sequence of Streptomyces sp. strain Wigar10, isolated from a surface-sterilized garlic bulb. *J. Bacteriol.*, **193**, 6999–7000.
22. Anand,S., Prasad,M.V., Yadav,G., Kumar,N., Shehara,J., Ansari,M.Z. and Mohanty,D. (2010) SBSPKS: structure based sequence analysis of polyketide synthases. *Nucleic Acids Res.*, **38**, W487–W496.
23. Medema,M.H., Blin,K., Cimermancic,P., de Jager,V., Zakrzewski,P., Fischbach,M.A., Weber,T., Takano,E. and Breitling,R. (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.*, **39**, W339–W346.
24. Starcevic,A., Zucko,J., Simunkovic,J., Long,P.F., Cullum,J. and Hranueli,D. (2008) ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. *Nucleic Acids Res.*, **36**, 6882–6892.
25. Kim,J. and Yi,G.S. (2012) PKMiner: a database for exploring type II polyketide synthases. *BMC Microbiol.*, **12**, 169.
26. Hopwood,D.A., Malpartida,F., Kieser,H.M., Ikeda,H., Duncan,J., Fujii,I., Rudd,B.A., Floss,H.G. and Omura,S. (1985) Production of 'hybrid' antibiotics by genetic engineering. *Nature*, **314**, 642–644.
27. McDaniel,R., Thamchaipenet,A., Gustafsson,C., Fu,H., Betlach,M. and Ashley,G. (1999) Multiple genetic modifications of the erythromycin polyketide synthase to produce a library of novel ''unnatural'' natural products. *Proc. Natl. Acad. Sci. USA*, **96**, 1846–1851.
28. Zhang,H., Wang,Y., Wu,J., Skalina,K. and Pfeifer,B.A. (2010) Complete biosynthesis of erythromycin A and designed analogs using E. coli as a heterologous host. *Chem. Biol.*, **17**, 1232–1240.
29. Horinouchi,S. (2009) Combinatorial biosynthesis of plant medicinal polyketides by microorganisms. *Curr. Opin. Chem. Biol.*, **13**, 197–204.
30. Watanabe,K. and Oikawa,H. (2007) Robust platform for de novo production of heterologous polyketides and nonribosomal peptides in Escherichia coli. *Org. Biomol. Chem.*, **5**, 593–602.
31. Komatsu,M., Uchiyama,T., Omura,S., Cane,D.E. and Ikeda,H. (2010) Genome-minimized Streptomyces host for the heterologous expression of secondary metabolism. *Proc. Natl. Acad. Sci. USA*, **107**, 2646–2651.
32. Khan,S.T., Komaki,H., Motohashi,K., Kozone,I., Mukai,A., Takagi,M. and Shin-ya,K. (2011) Streptomyces associated with a marine sponge Haliclona sp.; biosynthetic genes for secondary metabolites and products. *Environ. Microbiol.*, **13**, 391–403.
33. Komaki,H., Izumikawa,M., Ueda,J., Nakashima,T., Khan,S.T., Takagi,M. and Shin-ya,K. (2009) Discovery of a pimaricin analog JBIR-13, from Streptomyces bicolor NBRC 12746 as predicted by sequence analysis of type I polyketide synthase gene. *Appl. Microbiol. Biotechnol.*, **83**, 127–133.